An Automatic Finite-Sample Robustness Metric: Can Dropping a Little Data Make a Big Difference?

Ryan Giordano
MIT

Rachael Meager
LSE

Tamara Broderick
MIT

Job talk 2021

Dropping data: Motivation

You're a data analyst, and you've

- Gathered some exchangeable data,
- Cleaned up / removed outliers,
- Checked for correct specification, and
- Drawn a conclusion from your statistical analysis
  (e.g., based the sign / significance of some estimated parameter).

Dropping data: Motivation

You're a data analyst, and you've

- Gathered some exchangeable data,
- Cleaned up / removed outliers,
- Checked for correct specification, and
- Drawn a conclusion from your statistical analysis (e.g., based the sign / significance of some estimated parameter).

Well done!

Would you be concerned if you could **reverse your conclusion** by removing a **small proportion** (say, 0.1%) of your data?

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points.

The variable “Beta” estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original	0	-4.55 (5.88)

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points.

The variable “Beta” estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points.

The variable "Beta" estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)
Change significance	14	-10.96 (5.57)

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points.

The variable "Beta" estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)
Change significance	14	-10.96 (5.57)
Change both	15	7.03 (2.55)

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points.

The variable "Beta" estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)
Change significance	14	-10.96 (5.57)
Change both	15	7.03 (2.55)

By removing very few data points ($15/16560 \approx 0.1\%$), we can reverse the qualitative conclusions of the original study!

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points.

The variable “Beta” estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)
Change significance	14	-10.96 (5.57)
Change both	15	7.03 (2.55)

By removing very few data points ($15/16560 \approx 0.1\%$), we can reverse the qualitative conclusions of the original study!

Question: Is the reported interval $-4.55 \pm (5.88)$ a reasonable description of the uncertainty in the estimated efficacy of microcredit?

Dropping data: Motivation

Would you be concerned if you could **reverse your conclusion** by removing a **small proportion** (say, 0.1%) of your data?

Dropping data: Motivation

Would you be concerned if you could **reverse your conclusion** by removing a **small proportion** (say, 0.1%) of your data?

Not always!

Dropping data: Motivation

Would you be concerned if you could **reverse your conclusion** by removing a **small proportion** (say, 0.1%) of your data?

Not always!

...but sometimes, surely yes.

For example, often in economics:

- Small fractions of data are missing not-at-random,
- Policy population is different from analyzed population,
- We report a convenient summary (e.g. mean) of a complex effect,
- Models are stylized proxies of reality.

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Question 1: How do we find influential datapoints?

The number of subsets $\binom{N}{\lfloor \alpha N \rfloor}$ can be very large even when α is very small.

In the MX microcredit study, $\binom{16560}{15} \approx 1.4 \cdot 10^{51}$ sets to check for $\alpha = 0.0009$.

We provide a fast, automatic approximation based on the **influence function**.

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Question 1: How do we find influential datapoints?

Question 2: What makes an estimator non-robust?

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Question 1: How do we find influential datapoints?

Question 2: What makes an estimator non-robust?

Non-robustness to removal of $\lfloor \alpha N \rfloor$ points is:

- Not (necessarily) caused by misspecification.
- Not (necessarily) caused by outliers.
- Not captured by standard errors.
- Not mitigated by large N.
- Primarily determined by the **signal to noise** ratio

  ... in a sense which we will define.

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Question 1: How do we find influential datapoints?

Question 2: What makes an estimator non-robust?

Question 3: When is our approximation accurate?

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Question 1: How do we find influential datapoints?

Question 2: What makes an estimator non-robust?

Question 3: When is our approximation accurate?

- We provide deterministic error bounds for small α.
- We show the accuracy in simple experiments.
- We show the accuracy in a number of real-world experiments.

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Question 1: How do we find influential datapoints?

Question 2: What makes an estimator non-robust?

Question 3: When is our approximation accurate?

Conclusion: Related work and future directions

Question 1:
How do we find influential datapoints?

Which estimators do we study?

Suppose we have N data points $d_1, \ldots, d_N$. Then:

$$\hat{\theta} := \vec{\theta} \text{ such that } \sum_{n=1}^{N} G(\vec{\theta}, d_n) = 0_P.$$

Leave points out by setting their elements of $\vec{w}$ to zero.

These are "Z-estimators," i.e., roots of estimating equations.

Examples: all minimizers of empirical loss (OLS, MLE, VB), and more.

Which estimators do we study?

Suppose we have N data points $d_1, \ldots, d_N$. Then:

$$\hat{\theta}(\vec{w}) := \vec{\theta} \text{ such that } \sum_{n=1}^{N} \vec{w}_n G(\vec{\theta}, d_n) = 0_P.$$

Leave points out by setting their elements of $\vec{w}$ to zero.

These are "Z-estimators," i.e., roots of estimating equations.

Examples: all minimizers of empirical loss (OLS, MLE, VB), and more.

Which estimators do we study?

Suppose we have N data points $d_1, \ldots, d_N$. Then:

$$\hat{\theta}(\vec{w}) := \vec{\theta} \text{ such that } \sum_{n=1}^{N} \vec{w}_n G(\vec{\theta}, d_n) = 0_P.$$

Leave points out by setting their elements of $\vec{w}$ to zero.

Which estimators do we study?

Suppose we have N data points $d_1, \ldots, d_N$. Then:

$$\hat{\theta}(\vec{w}) := \vec{\theta} \ \text{ such that } \ \sum_{n=1}^{N} \vec{w}_n G(\vec{\theta}, d_n) = 0_P.$$

Leave points out by setting their elements of $\vec{w}$ to zero.

Fix a quantity of interest, $\phi(\vec{\theta})$:

$$\phi(\vec{\theta}) = \vec{\theta}_p$$

$$\phi(\vec{\theta}) = \vec{\theta}_p + \frac{1.96}{\sqrt{N}} \hat{\sigma}_\phi(\vec{\theta})$$

Which estimators do we study?

Suppose we have N data points $d_1, \ldots, d_N$. Then:

$$\hat{\theta}(\vec{w}) := \vec{\theta} \text{ such that } \sum_{n=1}^{N} \vec{w}_n G(\vec{\theta}, d_n) = 0_P.$$

Leave points out by setting their elements of $\vec{w}$ to zero.

Fix a quantity of interest, $\phi(\vec{\theta})$:

$$\phi(\vec{\theta}) = \vec{\theta}_p$$

$$\phi(\vec{\theta}) = \vec{\theta}_p + \frac{1.96}{\sqrt{N}} \hat{\sigma}_\phi(\vec{\theta})$$

Let the **"signal"**, Δ, be a "large" change in ϕ.

Which estimators do we study?

Suppose we have N data points $d_1, \ldots, d_N$. Then:

$$\hat{\theta}(\vec{w}) := \vec{\theta} \ \text{ such that } \ \sum_{n=1}^{N} \vec{w}_n G(\vec{\theta}, d_n) = 0_P.$$

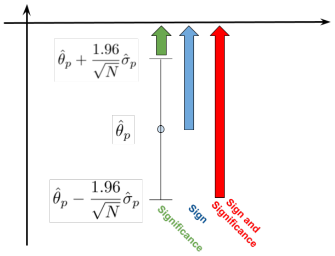
Leave points out by setting their elements of $\vec{w}$ to zero.

Fix a quantity of interest, $\phi(\vec{\theta})$:

$$\phi(\vec{\theta}) = \vec{\theta}_p$$

$$\phi(\vec{\theta}) = \vec{\theta}_p + \frac{1.96}{\sqrt{N}} \hat{\sigma}_\phi(\vec{\theta})$$

Let the **"signal"**, Δ, be a "large" change in ϕ.



Which estimators do we study?

Suppose we have N data points $d_1, \ldots, d_N$. Then:

$$\hat{\theta}(\vec{w}) := \vec{\theta} \text{ such that } \sum_{n=1}^{N} \vec{w}_n G(\vec{\theta}, d_n) = 0_P.$$

Leave points out by setting their elements of $\vec{w}$ to zero.

Fix a quantity of interest, $\phi(\vec{\theta})$.

Let the **"signal"**, Δ, be a "large" change in ϕ.

Can we reverse our conclusion by dropping $\lfloor \alpha N \rfloor$ datapoints?

Which estimators do we study?

Suppose we have N data points $d_1, \ldots, d_N$. Then:

$$\hat{\theta}(\vec{w}) := \vec{\theta} \text{ such that } \sum_{n=1}^{N} \vec{w}_n G(\vec{\theta}, d_n) = 0_P.$$

Leave points out by setting their elements of $\vec{w}$ to zero.

Fix a quantity of interest, $\phi(\vec{\theta})$.

Let the **"signal"**, Δ, be a "large" change in ϕ.

Can we reverse our conclusion by dropping $\lfloor \alpha N \rfloor$ datapoints? $\Leftrightarrow$

Is there a $\vec{w}$, with $\lfloor \alpha N \rfloor$ zeros, such that $\phi(\hat{\theta}(\vec{w})) - \phi(\hat{\theta}) \geq \Delta$?

Hard! Evaluating $\hat{\theta}(\vec{w})$ is costly and lots of $\vec{w}$ have $\lfloor \alpha N \rfloor$ zeros.

Taylor series approximation.

Is there a $\vec{w}$, with $\lfloor \alpha N \rfloor$ zeros, such that $\phi(\hat{\theta}(\vec{w})) - \phi(\hat{\theta}) \geq \Delta$?

To simplify the search over $\vec{w}$, we form the Taylor series approximation:

$$\phi(\hat{\theta}(\vec{w})) - \phi(\hat{\theta}) \approx \phi^{\mathrm{lin}}(\vec{w}) - \phi(\hat{\theta}) := - \sum_{n : \vec{w}_n = 0} \psi_n, \text{ where } \psi_n := \left. \frac{\partial \phi(\hat{\theta}(\vec{w}))}{\partial \vec{w}_n} \right|_{\vec{1}}.$$

Taylor series approximation.

Is there a $\vec{w}$, with $\lfloor \alpha N \rfloor$ zeros, such that $\phi(\hat{\theta}(\vec{w})) - \phi(\hat{\theta}) \geq \Delta$?

To simplify the search over $\vec{w}$, we form the Taylor series approximation:

$$\phi(\hat{\theta}(\vec{w})) - \phi(\hat{\theta}) \approx \phi^{\mathrm{lin}}(\vec{w}) - \phi(\hat{\theta}) := - \sum_{n: \vec{w}_n = 0} \psi_n, \text{ where } \psi_n := \left. \frac{\partial \phi(\hat{\theta}(\vec{w}))}{\partial \vec{w}_n} \right|_{\vec{1}}.$$

The values ψ_n are the **"empirical influence function."** [Hampel, 1986]

The ψ_n can be **easily and automatically** computed from $\hat{\theta}$.

The approximation is **typically accurate** for small α. [Giordano et al., 2019]

Taylor series approximation.

Is there a $\vec{w}$, with $\lfloor \alpha N \rfloor$ zeros, such that $\phi(\hat{\theta}(\vec{w})) - \phi(\hat{\theta}) \geq \Delta$?

To simplify the search over $\vec{w}$, we form the Taylor series approximation:

$$\phi(\hat{\theta}(\vec{w})) - \phi(\hat{\theta}) \approx \phi^{\mathrm{lin}}(\vec{w}) - \phi(\hat{\theta}) := -\sum_{n:\vec{w}_n=0} \psi_n, \text{ where } \psi_n := \left.\frac{\partial \phi(\hat{\theta}(\vec{w}))}{\partial \vec{w}_n}\right|_{\vec{1}}.$$

The values ψ_n are the **"empirical influence function."** [Hampel, 1986]

The ψ_n can be **easily and automatically** computed from $\hat{\theta}$.

The approximation is **typically accurate** for small α. [Giordano et al., 2019]

Is there a $\vec{w}$, with $\lfloor \alpha N \rfloor$ zeros, such that $\phi^{\mathrm{lin}}(\vec{w}) - \phi(\hat{\theta}) \geq \Delta$?

Taylor series approximation.

Is there a $\vec{w}$, with $\lfloor \alpha N \rfloor$ zeros, such that $\phi(\hat{\theta}(\vec{w})) - \phi(\hat{\theta}) \ge \Delta$?

To simplify the search over $\vec{w}$, we form the Taylor series approximation:

$$\phi(\hat{\theta}(\vec{w})) - \phi(\hat{\theta}) \approx \phi^{\mathrm{lin}}(\vec{w}) - \phi(\hat{\theta}) := - \sum_{n : \vec{w}_n = 0} \psi_n, \text{ where } \psi_n := \left. \frac{\partial \phi(\hat{\theta}(\vec{w}))}{\partial \vec{w}_n} \right|_{\vec{1}}.$$

The values ψ_n are the **"empirical influence function."** [Hampel, 1986]

The ψ_n can be **easily and automatically** computed from $\hat{\theta}$.

The approximation is **typically accurate** for small α. [Giordano et al., 2019]

Is there a $\vec{w}$, with $\lfloor \alpha N \rfloor$ zeros, such that $\phi^{\mathrm{lin}}(\vec{w}) - \phi(\hat{\theta}) \ge \Delta$?

Easy! The most influential points for $\phi^{\mathrm{lin}}(\vec{w})$ have the most negative ψ_n.

Taylor series approximation.

Procedure:

Taylor series approximation.

Procedure:

1. Compute the "original" estimator, $\hat{\theta}$ and $\phi(\hat{\theta})$.

Taylor series approximation.

Procedure:

1. Compute the "original" estimator, $\hat{\theta}$ and $\phi(\hat{\theta})$.
2. Compute and sort the influence scores, $\psi_{(1)} \leq \psi_{(2)} \leq \ldots \leq \psi_{(N)}$.

Taylor series approximation.

Procedure:

1. Compute the "original" estimator, $\hat{\theta}$ and $\phi(\hat{\theta})$.
2. Compute and sort the influence scores, $\psi_{(1)} \leq \psi_{(2)} \leq \ldots \leq \psi_{(N)}$.
3. Let $\vec{w}^*$ leave out the data corresponding to $\psi_{(1)}, \ldots, \psi_{(\lfloor \alpha N \rfloor)}$.

Taylor series approximation.

Procedure:

1. Compute the "original" estimator, $\hat{\theta}$ and $\phi(\hat{\theta})$.
2. Compute and sort the influence scores, $\psi_{(1)} \leq \psi_{(2)} \leq \ldots \leq \psi_{(N)}$.
3. Let $\vec{w}^*$ leave out the data corresponding to $\psi_{(1)}, \ldots, \psi_{(\lfloor \alpha N \rfloor)}$.
4. Report non-robustness if $\Delta \leq \phi^{\mathrm{lin}}(\vec{w}^*) - \phi(\hat{\theta}) = -\sum_{n=1}^{\lfloor \alpha N \rfloor} \psi_{(n)}$.

Taylor series approximation.

Procedure:

1. Compute the "original" estimator, $\hat{\theta}$ and $\phi(\hat{\theta})$.
2. Compute and sort the influence scores, $\psi_{(1)} \leq \psi_{(2)} \leq \ldots \leq \psi_{(N)}$.
3. Let $\vec{w}^*$ leave out the data corresponding to $\psi_{(1)}, \ldots, \psi_{(\lfloor \alpha N \rfloor)}$.
4. Report non-robustness if $\Delta \leq \phi^{\mathrm{lin}}(\vec{w}^*) - \phi(\hat{\theta}) = -\sum_{n=1}^{\lfloor \alpha N \rfloor} \psi_{(n)}$.

5 **Optional:** Compute $\hat{\theta}(\vec{w}^*)$, and verify that $\Delta \leq \phi(\hat{\theta}(\vec{w}^*)) - \phi(\hat{\theta})$.

Computing the influence function.

How to compute $\psi_n := \frac{\partial \phi(\hat{\theta}(\vec{w}))}{\partial \vec{w}_n}\Big|_{\vec{1}}$? Recall $\sum_{n=1}^{N} \vec{w}_n G(\hat{\theta}(\vec{w}), d_n) = 0_P$.

Step zero: Implement software to compute $G(\theta, d_n)$ and $\phi(\theta)$. Find $\hat{\theta}$.

Step one: By the chain rule, $\psi_n = \frac{\partial \phi(\hat{\theta}(\vec{w}))}{\partial \vec{w}_n}\Big|_{\vec{1}} = \frac{\partial \phi(\theta)}{\partial \theta^T}\Big|_{\hat{\theta}} \frac{\partial \hat{\theta}(\vec{w})}{\partial \vec{w}_n}\Big|_{\vec{1}}$.

Step two: By the implicit function theorem:

$$\frac{\partial \hat{\theta}(\vec{w})}{\partial \vec{w}_n}\Bigg|_{\vec{1}} = \frac{1}{N}\left(\frac{1}{N}\sum_{n'=1}^{N} \frac{\partial}{\partial \theta^T} G(\vec{\theta}, d_{n'})\Big|_{\hat{\theta}}\right)^{-1} G(\hat{\theta}, d_n).$$

Step three: Use *automatic differentiation* on $\phi(\theta)$ and $G(\theta, d_n)$ from step zero to compute $\frac{\partial \phi(\theta)}{\partial \theta^T}$ and $\frac{\partial}{\partial \theta^T} G(\vec{\theta}, d_n)$.

- The user does step zero. The rest is automatic.
- The primary computational expense is the Hessian inverse.
- Automatic differentiation is the chain rule applied to a program.
- Typically $\psi_n = O(N^{-1})$.

Question 2:
What makes an estimator non-robust?

What makes an estimator non-robust? A tail sum.

$$
\begin{aligned}
\Delta &\le \phi^{\mathrm{lin}}(\vec{w}^*) - \phi(\hat\theta) && \text{Report non-robustness} \\
&= -\sum_{n=1}^{\lfloor \alpha N \rfloor} \psi_{(n)} && \text{(By definition)} \\
&= -\frac{1}{N}\sum_{n=1}^{\lfloor \alpha N \rfloor} N\psi_{(n)} && \text{(Recall } \psi_n = O_p(N^{-1})\text{)} \\
&\le \underbrace{\left(\frac{1}{N}\sum_{n=1}^{N} N^2 \psi_{(n)}^2\right)^{1/2}}_{=:\, \hat\sigma_\phi} \underbrace{\left(\frac{1}{N}\sum_{n=1}^{N} \mathbb{I}\left(n \le \lfloor \alpha N \rfloor\right)\right)^{1/2}}_{=:\, \mathcal{S}_\alpha \le \sqrt{\alpha}} && \text{(Cauchy-Schwartz)}
\end{aligned}
$$

Suppose that $\hat\theta \xrightarrow{p} \theta_0$ and $\phi(\hat\theta) \rightsquigarrow \mathcal{N}(\phi(\theta_0), \sigma^2)$.

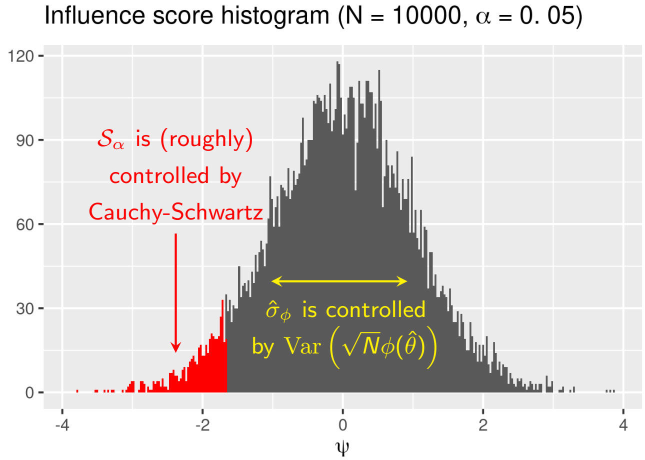
Typically, $\hat\sigma_\phi \xrightarrow{p} \sigma$ [Hampel, 1986].

A slightly more careful analysis gives $\mathcal{S}_\alpha \le \sqrt{\alpha(1-\alpha)}$.

What makes an estimator non-robust? A tail sum.

Report non-robustness if the **"signal to noise ratio"** $\frac{\Delta}{\hat{\sigma}_\phi} \le \mathcal{S}_\alpha$ where

- The "noise" $\hat{\sigma}_\phi^2 \to \mathrm{Var}(\sqrt{N}\phi)$ [Hampel, 1986]
- The "shape" $\mathcal{S}_\alpha \le \sqrt{\alpha(1-\alpha)}$ and converges to a nonzero constant



Corollaries.

Corollaries.

Report non-robustness if the **“signal to noise ratio”** $\frac{\Delta}{\hat{\sigma}_\phi} \leq \mathcal{S}_\alpha$.

Corollaries.

Report non-robustness if the **"signal to noise ratio"** $\frac{\Delta}{\hat{\sigma}_\phi} \leq \mathcal{S}_\alpha$.

Corollary: Non-robustness possible even with correct specification.

Corollaries.

Report non-robustness if the **"signal to noise ratio"** $\frac{\Delta}{\hat{\sigma}_\phi} \leq \mathcal{S}_\alpha$.

Corollary: Non-robustness possible even with correct specification.

Corollary: Leave-$\lfloor \alpha N \rfloor$-out robustness does not vanish as $N \to \infty$.

Corollaries.

Report non-robustness if the **"signal to noise ratio"** $\frac{\Delta}{\hat{\sigma}_\phi} \le \mathcal{S}_\alpha$.

Corollary: Non-robustness possible even with correct specification.

Corollary: Leave-$\lfloor \alpha N \rfloor$-out robustness does not vanish as $N \to \infty$.

Recall that standard errors reject when $\frac{\Delta}{\hat{\sigma}_\phi} \le \frac{1.96}{\sqrt{N}}$.

Corollaries.

Report non-robustness if the **"signal to noise ratio"** $\frac{\Delta}{\hat{\sigma}_\phi} \leq \mathcal{S}_\alpha$.

Corollary: Non-robustness possible even with correct specification.

Corollary: Leave-$\lfloor \alpha N \rfloor$-out robustness does not vanish as $N \to \infty$.

Recall that standard errors reject when $\frac{\Delta}{\hat{\sigma}_\phi} \leq \frac{1.96}{\sqrt{N}}$.

Corollary: Leave-$\lfloor \alpha N \rfloor$-out is different from standard errors.

Corollaries.

Report non-robustness if the **"signal to noise ratio"** $\frac{\Delta}{\hat{\sigma}_\phi} \leq \mathcal{S}_\alpha$.

Corollary: Non-robustness possible even with correct specification.

Corollary: Leave-$\lfloor \alpha N \rfloor$-out robustness does not vanish as $N \to \infty$.

Recall that standard errors reject when $\frac{\Delta}{\hat{\sigma}_\phi} \leq \frac{1.96}{\sqrt{N}}$.

Corollary: Leave-$\lfloor \alpha N \rfloor$-out is different from standard errors.

Corollary: Insignificance is always non-robust.
Take $\Delta = \frac{1.96\hat{\sigma}_\phi}{\sqrt{N}} \to 0 \leq \mathcal{S}_\alpha$.

Corollaries.

Report non-robustness if the **"signal to noise ratio"** $\frac{\Delta}{\hat{\sigma}_\phi} \leq \mathcal{S}_\alpha$.

Corollary: Non-robustness possible even with correct specification.

Corollary: Leave-$\lfloor \alpha N \rfloor$-out robustness does not vanish as $N \to \infty$.

Recall that standard errors reject when $\frac{\Delta}{\hat{\sigma}_\phi} \leq \frac{1.96}{\sqrt{N}}$.

Corollary: Leave-$\lfloor \alpha N \rfloor$-out is different from standard errors.

Corollary: Insignificance is always non-robust.
Take $\Delta = \frac{1.96 \hat{\sigma}_\phi}{\sqrt{N}} \to 0 \leq \mathcal{S}_\alpha$.

Corollary: Gross outliers primarily affect robustness through $\hat{\sigma}_\phi$.
Cauchy-Schwartz is tight when all the influence scores are the same.

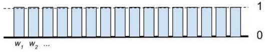
Question 3:
When is our approximation accurate?

The influence function

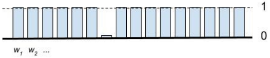
- Weights as derivatives
- Influence function
- Simulation
- Experiments

The linear approximation.

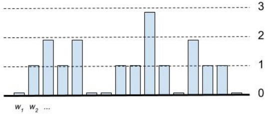
Original weights:

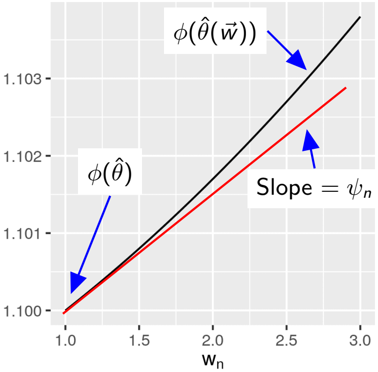


Leave-one-out weights:



Bootstrap weights:





$$\phi(\hat{\theta}(\vec{w})) = \phi(\hat{\theta}) + \sum_{n=1}^{N} \psi_n(\vec{w}_n - 1) + \text{Higher-order derivatives}$$

Key idea: Controlling higher-order derivatives can control the error.

The linear approximation.

Assumption ((?, Assumptions 1-4))

Let W_α be the set of weight vectors with no more than $\lfloor \alpha N \rfloor$ zeros as given by Eq. **??**. *Assume there exists a compact domain $\Omega_\theta \subseteq \mathbb{R}^D$ containing $\hat{\theta}(\vec{w})$ for all $\vec{w} \in W_\alpha$, such that*

1. *For all $\theta \in \Omega_\theta$ and all n, $\theta \mapsto G(\theta, d_n)$ is continuously differentiable with derivative*

$$\left. \frac{\partial G(\theta, d_n)}{\partial \theta^T} \right|_\theta =: H(\theta, d_n).$$

2. *For all $\theta \in \Omega_\theta$, there exists $C_{op} < \infty$ such that* $\sup_{\theta \in \Omega_\theta} \left\| \frac{1}{N} \sum_{n=1}^N H(\theta, d_n) \right\|_{op} \leq C_{op}$.
3. *There exists a constant $C_{gh} < \infty$ such that*

$$\sup_{\theta \in \Omega_\theta} \max \left\{ \frac{1}{N} \sum_{n=1}^N \| G(\theta, d_n) \|_2^2 , \frac{1}{N} \sum_{n=1}^N \| H(\theta, d_n) \|_2^2 \right\} \leq C_{gh}^2.$$

Conclusions

Conclusion

- You may be concerned if you could reverse your conclusion by removing a $\lfloor \alpha N \rfloor$ datapoints, for some small α.

Conclusion

- You may be concerned if you could reverse your conclusion by removing a $\lfloor \alpha N \rfloor$ datapoints, for some small α.
- Robustness to removing a $\lfloor \alpha N \rfloor$ datapoints is principally determined by the signal to noise ratio, does not disappear asymptotically, and is distinct from (and typically larger than) standard errors.

Conclusion

- You may be concerned if you could reverse your conclusion by removing a $\lfloor \alpha N \rfloor$ datapoints, for some small α.
- Robustness to removing a $\lfloor \alpha N \rfloor$ datapoints is principally determined by the signal to noise ratio, does not disappear asymptotically, and is distinct from (and typically larger than) standard errors.
- Robustness to removing a $\lfloor \alpha N \rfloor$ datapoints is easy to check! We can quickly and automatically find an approximate influential set which is accurate for small α.

Links and references

Tamara Broderick, Ryan Giordano, Rachael Meager (alphabetical authors)
"An Automatic Finite-Sample Robustness Metric: Can Dropping a Little Data Change Conclusions?"
https://arxiv.org/abs/2011.14999

See the paper for applications to:
- Hierarchical meta-analysis of microcredit [Meager, 2020]
- Cash transfers randomized controlled trial [Angelucci and De Giorgi, 2009]
- Oregon Medicaid experiment [Finkelstein et al., 2012]
- Expository simulations

`zaminfluence`: `R` package with leave-α-out robustness for OLS and IV estimators
https://github.com/rgiordan/zaminfluence

M. Angelucci and G. De Giorgi. Indirect effects of an aid program: How do cash transfers affect ineligibles' consumption? *American Economic Review*, 99(1):486–508, 2009.

M. Angelucci, D. Karlan, and J. Zinman. Microcredit impacts: Evidence from a randomized microcredit program placement experiment by Compartamos Banco. *American Economic Journal: Applied Economics*, 7(1):151–82, 2015.

A. Finkelstein, S. Taubman, B. Wright, M. Bernstein, J. Gruber, J. Newhouse, H. Allen, K. Baicker, and Oregon Health Study Group. The Oregon health insurance experiment: Evidence from the first year. *The Quarterly Journal of Economics*, 127(3):1057–1106, 2012.

R. Giordano, M. I. Jordan, and T. Broderick. A higher-order Swiss army infinitesimal jackknife. *arXiv preprint arXiv:1907.12116*, 2019.

F. Hampel. *Robust statistics: The approach based on influence functions*, volume 196. Wiley-Interscience, 1986.

R. Meager. Aggregating distributional treatment effects: A Bayesian hierarchical analysis of the microcredit literature. *LSE working paper*, 2020.